

United States Patent Application

for

METHOD AND SYSTEM FOR MEASURING THE QUALITY OF A
HIERARCHY

Inventors:

George H. Forman
Tom Fawcett
Henri J. Suermondt

METHOD AND SYSTEM FOR MEASURING THE QUALITY OF A HIERARCHY

FIELD OF THE INVENTION

The present invention relates generally to hierarchies, and more particularly, to a method for measuring the quality of a hierarchy.

BACKGROUND OF THE INVENTION

Portals (e.g., Yahoo) arrange Web sites into a topic hierarchy in order to facilitate and aid a user in finding web sites of interest. FIG. 6 illustrates a portion of an exemplary topic hierarchy. In this topic hierarchy, there is a topic entitled "Health" and a sibling topic entitled "Entertainment". The "Health" topic has two sub-topics (or children nodes): "Diseases" and "Doctors". The "Entertainment" topic has two sub-topics: "Soccer" and "Chess".

Another use of a topic hierarchy is to organize content on a particular Web site. For example, HP (the assignee of the present patent application) organizes its technical notes and publications in hierarchies for ease of browsing.

Hierarchies are typically designed in the following manner. First, a user generates topics or categories into which the content may be filed, including their hierarchical relationships to one another. Second, content (e.g., web sites or technical articles) is placed under appropriate topics in the hierarchy. For example, each document is filed under one of the topics. As new documents become available, these new documents must also be filed under one of the topics. When a document does not appear to fit into any of the current topics, the user can then add new topics to the hierarchy. Similarly, the user can delete topics or modify current topics in the hierarchy or their arrangement. It is noted that whenever topics are added, deleted, or otherwise modified, the user must then evaluate whether any of the documents in the hierarchy need to be re-classified to a different topic.

As can be appreciated, this process of placing new content into the hierarchy and of maintaining the topics in a hierarchy is labor intensive. One can envision cases where it is not practical for human agents to perform the categorization of new content into the hierarchy because of the sheer volume of the documents or web sites that require categorization.

Some have suggested and attempted to utilize automated categorization programs that are based on text categorization technology from the field of artificial intelligence to automate the process of placing new content into the hierarchy.

Automated categorization programs that are based on machine learning operate in the following manner. First, a hierarchy of topics is provided to the automated categorization program. Second, training examples are provided to the automated categorization program. These training examples train the program to classify new content in a manner similar to how the training examples are classified into predetermined topics. Some examples of such automated categorization programs include the well-known Naïve Bayes and C4.5 algorithms, as well as commercial offerings by companies such as Autonomy Inc.

Unfortunately, the quality of the categorization generated by automated categorization programs depends on how well the automated categorization programs can "interpret" the hierarchy. For example, topics or categories that are sensible to a human user may confuse an automated categorization computer program. The topics "Chess" and "Soccer" can reasonably be grouped under the parent topic "Entertainment." However, it may be difficult, if not impossible, for an automated categorization computer program to find common words or other text that would suggest that both sub-topics "Chess" and "Soccer" should be under the topic "Entertainment."

In this regard, it is desirable for there to be a mechanism that analyses hierarchies and determines the quality of the arrangement of topics and corresponding documents for each place (e.g., particular topic subtree) in the hierarchy. This mechanism facilitates the design of hierarchies in such a way as to tailor the designed hierarchies so

-4-

that automated categorization programs can place content therein in an efficient and accurate manner.

Based on the foregoing, there remains a need for a mechanism to determine a measure of coherence for the arrangement of hierarchically organized topics at each place in the hierarchy.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
22

SUMMARY OF THE INVENTION

One aspect of the present invention is the provision of a method to determine a measure of coherence for the arrangement of hierarchically organized topics at each place in the hierarchy.

Another aspect of the present invention is the use of this measure of hierarchical coherence to design hierarchies that are tailored for automated categorization of content therein is described.

Another aspect of the present invention is the provision of a mechanism for determining a measure of coherence for the arrangement of hierarchically organized topics at each place in the hierarchy based on the distribution of features in a plurality of training cases filed into the hierarchy.

According to one embodiment, a method for determining a measure of coherence for the arrangement of hierarchically organized topics at each place in the hierarchy based on the distribution of features in a plurality of training cases filed into the hierarchy is described. The method measures the degree of coherence of all nodes in a hierarchy except leaf nodes and the root node. A hierarchy that includes a plurality of nodes (e.g., topics and sub-topics) is received. A plurality of training cases (e.g., documents appropriately filed into the hierarchy) is also received.

The following computation may be performed at each node in the hierarchy, except the root and the leaves: Based on the hierarchy and the training cases, determine a list of the most predictive features (e.g. words) that distinguish documents of the current node's sub-tree from those in its "local environment" (defined as the sub-trees of the current node's siblings as well as the parent node itself, if the parent contains any training cases). Optionally, any predictive features that are not represented fairly uniformly among the children subtrees of the current node based on the training cases under each child subtopic is eliminated from the list. If the list contains no features, assign a coherence value to indicate no coherence. Otherwise, assign a coherence value to indicate a level of coherence that depends on either the list of predictive features, their

degree of predictiveness, their degree of prevalence, the degree of uniform prevalence among the node's subtopics, or a combination thereof.

Other features and advantages of the present invention will be apparent from the detailed description that follows.

[illegible]

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements.

5 FIG. 1 illustrates an environment in which a coherence analyzer of the present invention may be implemented according to one embodiment of the present invention.

FIG. 2 is a block diagram illustrating in greater detail the coherence analyzer of FIG. 1.

10 FIG. 3 is a flow chart illustrating the processing steps performed by the coherence analyzer of FIG. 1 in accordance with one embodiment of the present invention.

FIG. 4 illustrates an exemplary hierarchy.

FIG. 5 illustrates an exemplary hierarchy with coherence measures assigned to each non-leaf node.

15 FIG. 6 illustrates a portion of an exemplary topic hierarchy.

DETAILED DESCRIPTION

A method for determining a measure of coherence for the arrangement of hierarchically organized topics at each place in the hierarchy. This measure is referred to herein as "hierarchical coherence" or simply "coherence." is described. In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present invention.

The following notation is utilized herein. The notation "D[^]" refers to the entire subtree rooted at the topic/directory D. The notation "D@" refers to the directory D only, excluding its children/descendants.

Environment for Coherence Analyzer 110

FIG. 1 illustrates an environment 100 in which a coherence analyzer 110 of the present invention may be implemented according to one embodiment of the present invention. The environment 100 includes a feature extractor 124, a coherence analyzer 110, and a user-interface presentation unit 150. The feature extractor 124 generates a set of labeled feature vectors 128, which can be, for example, training cases, based on a set of labeled documents (hereinafter referred to also as training cases) or feature guidelines 130. As used herein, the term "labeled" indicates that each training case, feature vector, or document is annotated with a node of the hierarchy where it should be filed. It is noted that the feature extractor 124 is needed for text domains. However, the feature extractor 124 may not be included for other domains where the training items contain a pre-prepared vector of features, such as in categorizing terrain types in satellite images by the values of neighboring pixels, or in recognizing postal zip code digits where the input data has already been converted to a feature vector. The user-interface

presentation unit 150 receives the coherence metric numbers 144 from the coherence analyzer 110 and generates a graphical display of the same for viewing by a user. It may, for example, sort the nodes by the assigned coherence metric to present the user with a list of the most or least coherent nodes.

5 The coherence analyzer 110 includes the following inputs. The coherence analyzer 110 includes a first input for receiving a hierarchy of topics 114 and a second input for receiving a set of labeled feature vectors 128. Based on these inputs, the coherence analyzer 110 generates a measure of coherence 144 for the arrangement of hierarchically organized topics at each place in the hierarchy (e.g., coherence metric
10 numbers).

Examples of a hierarchy of topics 114 include, but are not limited to, a directory hierarchy or email folder hierarchy. An example of training cases 118 that are filed under the topics include documents, such as text files or Web pages in directories, or emails in folders. It is noted that training cases 118 as described hereinafter with
15 reference to embodiments of the present invention refer to documents. However, training cases 118 can include any type of training case or training example.

Features

20 In situations where the training cases 118 have not previously been reduced to a set 128 of features, a standard and necessary pre-processing step to the coherence analyzer 110 includes a feature extractor 124 for decomposing each document into a set 128 of features. The set 128 of features can be, for example, the individual words of each document. In one embodiment, guidelines 130 may be provided to the feature extractor 124, and the feature extractor 124 generates a set 128 of features based on the
25 guidelines. A user may program these guidelines 130. For example, the guidelines may specify that words are to be considered any consecutive sequence of alphanumeric characters that are forced to lowercase. Furthermore, the guidelines may specify a common “bag of words” model, selecting those words that occur in less than twenty-five

percent (25%) of the documents, and that occur in more than twenty-five (25) documents overall). In another embodiment, in lieu of the previously described guidelines 130, a set of feature definitions (e.g., a given list of words to search for) is provided to the feature extractor 124.

5 A feature may be anything measurable about the document or training example. For example, in a hierarchy of foods a feature may be the percentage of USDA daily allowance of Vitamin B12 or grams of saturated fat.

In a hierarchy of documents, a set of features may be the individual words (e.g., single words and 2-word phrases) that occur in the set of documents, as with the standard “bag of words” model. In the preferred embodiment, the set of features 10 includes Boolean indicators of the presence or absence of each word that appears in the training set, except those words that occur in greater than a predetermined percentage of documents and except those words that occur in less than a predetermined number of occurrences. By excluding the words that occur greater than a predetermined percentage 15 (e.g., twenty percent) of all the documents, stopwords (e.g., “the” and “a”), which do not contribute to the coherence measure, are avoided. Similarly, rare words, such as those words that occur less than a predetermined number of times (e.g., 20 times overall) are excluded, since these words do not affect the coherence measure.

It is noted that a wide variety of feature engineering and feature selection 20 strategies, known to those skilled in the art, may be employed to determine the set of features. For example, feature engineering may look for 2-word phrases or 3-word phrases or restrict attention to noun phrases only. Features may also be negated to create new features, for example, the a Boolean indicator whose “true” value indicates the absence of the word “fun” may be strongly predictive for the “Health” category. 25 Other features can include, but are not limited to, document length, file extension type, or anything else about a document. Feature selection techniques can include selecting only those features with the highest “information gain” or “mutual information” metrics, as described in standard machine learning textbooks. Other feature engineering and

feature selection strategies that are known to those of ordinary skill in the art may also be applied in determining a set of features for use with the training examples (e.g., documents).

5 Coherence Metric

The coherence analyzer 110 assigns a coherence metric number 144 for each place (e.g., node) in the hierarchy, except the root and leaves. The coherence measure or metric 144 can be any value in the range 0% to 100%, with 0% indicating no coherence and 100% indicating complete coherence. Values will typically fall between 20% and 10 80%.

The coherence measure 144 is an indicator of how "natural" the grouping of subtopics under a node is, with respect to the topics beside and immediately above that topic (i.e., whether the documents under the current topic's subtrees have shared features that distinguish them as a whole from the documents in its "local environment" (defined 15 as the documents within sibling topics and documents assigned to the immediate parent). The coherence metric is not computed for the root node (which has no local environment) or for leaf nodes (which have no subtopics).

For example, referring to FIG. 4, if the word feature "medicine" appears in 100% of the documents at or under the topic "Health" and does not occur very often in the 20 documents under the topic "Entertainment", then the node "Health" would receive a hierarchical coherence of 1.0. Suppose that the only predictive feature for "Entertainment" is the word "fun" and that it appears in 60% of the documents under "Entertainment" and only very rarely under "Health." If the word "fun" occurs only under the subtopic "Soccer" and not under the subtopic "Chess" (i.e. non-uniform over 25 subtopics), then the "Entertainment" node will have a low coherence (e.g., a coherence value (CV) of 0%). On the other hand, if the word "fun" occurs with roughly the same prevalence under both "Chess" and "Soccer" (uniformity), then the "Entertainment" node receives a hierarchical coherence of 60%.

Coherence Analyzer 110

FIG. 2 is a block diagram illustrating in greater detail the coherence analyzer 110 of FIG. 1. The coherence analyzer 110 further includes a training case counter 210 for determining the number 214 of training cases (e.g., documents in each subtree). The coherence analyzer 110 further includes an average prevalence determination unit 220 for determining each feature's average prevalence 224 (i.e., average value in the documents in the subtree). For example, determining that the word "chess" appears in 95% of the documents in a particular subtree.

The coherence analyzer 110 further includes a predictive feature determination unit 230 for determining a set of predictive features 234 under each topic, optionally annotated with a number indicating their degree of predictiveness. Specifically, the predictive feature determination unit 230 determines the individual features that are most predictive of the entire subtree rooted at the topic or directory D (referred to herein as D^{\wedge}) as compared with its sibling subtrees or its parent node. Predictive features 234 are those features whose presence indicates a much higher probability that the document belongs in the D^{\wedge} subtree instead of in D's sibling subtrees or in D's parent node. A preferred method for generating predictive features 234 is described in greater detail hereinafter with reference to FIG. 3.

The coherence analyzer 110 further includes a subtopic uniformity determination unit 240 for determining which of the predictive features determined previously are also uniformly common among the subtrees and for each topic. The subtopic uniformity determination unit 240 generates a list of uniform predictive features 244 that may include a number to indicate their degree of uniformity. The coherence analyzer 110 also includes a coherence assignment unit 250 for generating a coherence measure 144 (e.g., a coherence metric number) based on a list of predictive features.

In one embodiment, the assignment of a coherence value to a current node is based on the list of predictive features, their degree of predictiveness, their degree of

prevalence, their degree of uniformity, or a combination thereof. It is noted that the degree of uniformity reflects how evenly distributed the predictive features are among the children subtrees of the current node based on the training cases under each child subtree. A preferred method for generating a coherence measure is described in greater detail hereinafter with reference to FIG. 3.

Processing Steps

FIG. 3 is a flow chart illustrating the processing steps performed by the coherence analyzer of FIG. 1 and FIG 2 in accordance with one embodiment of the present invention. In step 304, a hierarchy (e.g., a topic hierarchy) and a set of labeled training cases is received. The hierarchy is comprised of a plurality of nodes arranged in a tree. The plurality of nodes has at least one node under consideration (NUC). Each node under consideration has associated therewith its subtree and its "local environment" (i.e., its parent and the subtrees of its siblings), which is described hereinafter with reference to FIG. 4. The set of labeled training cases can be either documents or feature vectors. By "labeled" we mean that each training case is filed under a node of the hierarchy. If the training cases are documents (as opposed to feature vectors), each document is converted into a feature vector in processing step 308, which is referred to as feature extraction.

In step 310, the number of training cases (e.g., documents) under each topic subtree is determined. In step 320, the average prevalence (AP) for each feature under each topic subtree is determined (e.g., determining that the word feature "ball" appears in 90% of the documents under Soccer⁴).

In step 330, it is determined which features are predictive for each subtree versus the environment of the node under consideration based on the average prevalence and on the number of training cases. In a preferred embodiment, a statistical test, known as Fisher's Exact Test, is utilized. The Fisher's Exact Test provides more sensible results than Chi-Squared when the number of documents is small. To select a variable length

set of the "most" predictive words, a probability threshold of, for example, 0.001 is utilized against the output of Fisher's Exact Test.

Alternative strategies for selecting the most predictive features (e.g., words) include employing metrics, such as lift, odds-ratio, information-gain, and Chi-Squared.

- 5 As for selecting the "most" predictive, instead of selecting all those above some threshold, one might select the top 50 words or dynamically select the threshold. Other strategies that are known to those of ordinary skill in the art may also be utilized to select the most predictive words.

- 10 In step 334, it is determined which features that were selected in step 330 are also "uniformly common" among the subtrees. For example, the uniform predictive features for a topic are determined based on the average prevalence and the number of training cases under each of the subtrees of the topic. It is noted that in some embodiments, step 334 may be entirely absent.

- 15 In a preferred embodiment, whether a feature is "uniformly common" among the subtrees is determined by a "cosine similarity" test between the number of documents in each of the children subtrees and the feature occurrence counts in the subtrees. Those features with a cosine similarity greater than or equal to a threshold θ (in the preferred embodiment, we set θ to 0.90) are selected. Mathematically, features that meet the following criterion are selected:

$$\frac{\text{dotproduct}(F, N)}{\text{length}(F) * \text{length}(N)} \geq \theta$$

- 25 where F is a vector representing the feature occurrence counts for each child subtree (from step 320), and N is a vector representing the number of documents for each child subtree (from step 310). An array of features that are sorted by this metric may be stored.

Other strategies known in the art for selecting features that are "uniformly common" include selecting those features whose average prevalence feature vectors have the greatest projection along the distribution vector among the children subtopics of D, or selecting features that most likely fit the null hypothesis of the Chi Squared test.

5 In step 338, for each directory D in the hierarchy, except the root and the leaves, a hierarchical coherence number is generated and provided as output.

It is noted that assigning a coherence value to the current node indicating the current node's level of coherence may be based on one or more of the following: a list of predictive features, the degree of predictiveness of the predictive features, the degree of prevalence of the predictive features, and the degree of uniformity of the predictive features among the current node's subtopics. The degree of prevalence in X^\wedge indicates how frequently the word appears in documents under node X^\wedge . The degree of uniformity indicates how uniformly a word appears in each of X^\wedge 's subtopics, regardless of how prevalent the word is overall. It is noted that a feature that is deemed predictive does not automatically mean the feature is prevalent or uniform. For example, a feature may be predictive because it appears in 10% of X^\wedge documents and in 0% of documents in X^\wedge 's local environment (i.e. not highly prevalent) and may appear in only one of X^\wedge 's subtopics (i.e. not uniform).

20 In one embodiment, a coherence value is assigned to a particular topic or directory based on the average prevalence of one or more predictive and uniformly common features in step 338. In this embodiment, the hierarchical coherence of directory D may be defined as the overall prevalence of those features selected previously. When no features are selected, then the hierarchical coherence number for directory D is assigned a zero value.

25 For example, the feature having the greatest cosine similarity (e.g., $S[0]$ from the previous step) is selected, and the hierarchical coherence number is assigned the feature's average prevalence (from step 320) for the whole subtree D^\wedge .

In a preferred embodiment, the hierarchical coherence number is assigned an exponentially weighted average value over the most uniform features selected in the previous step. In other words, for the i -th feature [$i=0..$] from the sorted list recorded previously, a weighted average is computed of the feature average prevalence values (from step 320) using a weight of e^{-i} (i.e. the following schedule of weights is used: 64%, 23%, 9%, 3%, 1%). Because of the exponential fall-off, all remaining terms yield a fairly insignificant effect, and consequently, may be ignored. A weighted average value (e.g., an exponentially weighted average value) is utilized in this embodiment since there are some cases where it is not desirable for the metric to be dependent on a single feature alone. Moreover, a weighted average value prevents the metric from being overly sensitive to which individual features are selected in the feature extraction (step 124). Another reason for using a weighted average value is that certain features may have noise (e.g., the authors of a document may use synonyms for a concept). Other strategies include simply taking the average value of the top k features ($k = 1, 2, 3$, etc.) or using other weighting schedules, such as $1/i$.

Alternately, the determination of hierarchical coherence of step 338 may employ the maximum weighted projection of any feature selected in step 330. In another alternative embodiment, the determination of hierarchical coherence of step 338 employs the maximum average prevalence of any feature selected in step 330. In another alternative embodiment, the prevalence of each feature may be reduced by some degree based on how non-uniformly the feature is present in the child subtopics.

In another embodiment, there may be a post-processing step that outputs at each node D a mathematical aggregation function (e.g. sum, average, weighted-average, minimum, and maximum) of the coherence values that have been computed for its children nodes, thereby providing a measure of aggregated coherence that directly predicts the difficulty of choosing the correct subtree for a known-manner top-down or "Pachinko" classifier. With this extension to the method, a node that has many incoherent children has a low aggregate coherence value, suggesting a location in the

hierarchy where a Pachinko classifier is likely to make many errors and/or need additional training examples. Under this post-processing step, the root is assigned an aggregate coherence value, and there is no aggregate coherence value for nodes whose children are all leaves.

FIG. 5 illustrates an exemplary hierarchy with coherence measures assigned to each non-leaf node. The hierarchy 500 includes a root node 504, a current node 520, and a parent node 510 of the current node 520. The current node 520 includes a plurality of documents 528 or training cases and a coherence value (CV) 524. The current node 520 can have one or more sibling nodes 530, where each sibling node may have a corresponding sub-tree.

The current node 520 includes a subtree 550 that includes child nodes 538 and may include one or more leaf nodes 540. The subtree 550 is rooted at the current node. The coherence value 524 is an indicator of the existence of features (e.g., a keyword) that is common to the documents in the sub-tree 550 of the current node and yet distinguishes (e.g., uncommon) from the documents of the local environment 560 (i.e., documents in the siblings' subtrees and the documents in the parent node 510). The coherence analyzer 110 of the present invention generates a coherence value (CV) for each node in the hierarchy 500 except for leaf nodes and the root node.

It is noted that the predictiveness or a measure thereof may be determined by the training cases (e.g., documents) in the local environment 560 and the training cases in the subtree 550 of the current node.

Exemplary Applications

Some applications where the coherence analysis method of present invention may be applied include the organization of a set of hierarchical folders into which electronic mail may be sorted. An electronic mail software package, such as Microsoft Outlook or IBM Notes, may incorporate an automatic facility to categorize incoming electronic mail into hierarchical folders; such categorization may be improved by

performing the coherence analysis method on the collection of folders periodically, and improving the organization of the hierarchy based on the results.

Another application for the coherence analysis method of present invention is in the organization of a topic hierarchy at a news service. Based on the results, incoming news articles (e.g., Reuters & AP articles) may be automatically categorized with greater accuracy into a topic hierarchy at news web sites such as CNN.com.

Yet another application for the coherence analysis method of present invention is in the organization of a directory hierarchy at a search engine website. For example, a Web crawler automatically inserts entries into the Excite or AltaVista directory hierarchies.

Yet another application for the coherence analysis method of present invention is a hierarchy of new products at a portal, such as Yahoo Shopping or UDDI hierarchical business directories.

In summary, the coherence analysis method of present invention may be useful in any scenario where statistical or machine learning techniques are utilized to automatically categorize items into a hierarchy.

As can be appreciated, the maintainers of any of the above applications desire the highest achievable accuracy by the categorizer. It is noted that mis-located documents are generally annoying and costly. The training and accuracy of an automated top-down classifier trained by machine learning (e.g. Pachinko machine classifier) is likely to perform better when the hierarchy is coherent (i.e., there are features or words that characterize whole subtrees). The present invention provides the maintainer a way to measure the hierarchical coherence at each node, thereby identifying the least coherent subtrees.

Once a coherence measure is assigned to each node of the hierarchy by the present invention, maintainers can utilize this information to re-arrange the hierarchy to be more coherent, thereby leading to greater accuracy by the categorization technology. Alternatively, the coherence measure may indicate certain nodes or topics or sub-topics,

where more training examples added thereto may be needed to improve the performance of the classifier. In another scenario, the coherence measure may be utilized to choose or apply a particular technology to classify a particular portion of the hierarchy (e.g., sub-trees). In this manner, a fast, but less powerful classifier may be utilized to classify
5 for those nodes that have a high coherence value. A slower, but more powerful classifier or classifying technology is employed to classify documents into those sub-trees with nodes with low coherence measure. In this manner, the classification may be performed in an efficient manner, and resources are intelligently selected to suit a particular task at hand.

10 Alternatively, places in the hierarchy exhibiting poor coherence may be dealt with by modifying the classifier's structure (e.g., by deviating from the given hierarchy only for the purpose of more accurate classification).

For example, referring to FIG. 4, suppose that the node Entertainment exhibited low coherence. For the purpose of top-down classification only, the children subtopics, Soccer and Chess, may be moved so that they attach directly to the parent of
15 Entertainment. Alternately, supposing that the topic Entertainment contained many subtopics, and through a guessing or systematic search process, it is determined that eliminating the subtopic Chess greatly improves the coherence of topic Entertainment. Consequently, the subtopic Chess can be moved to be a sibling of Entertainment for the
20 purpose of improving top-down classification accuracy.

In the foregoing specification, the invention has been described with reference to specific embodiments thereof. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader scope of the invention. The specification and drawings are, accordingly, to be regarded in an
25 illustrative rather than a restrictive sense.